Getting Technical with TV&C...

Volume 1, Issue 1

Fall 1999

Standard Scores

What is a Standard Score?

A standard score is a linear (or in some cases a nonlinear) transformation of an original raw score. Standard scores are the result of a mathematical manipulation used to assign a common meaning to raw scores which have been derived from different scales or different units of measurement. In the case of a linear transformation, standard scores retain the exact numerical relationship to the original raw scores, because they are computed by merely subtracting one constant from each raw score (the mean) and then dividing the result by another constant (the standard deviation). The relative magnitude of differences between standard scores derived by a linear transformation corresponds exactly to the magnitude of the raw score differences. All properties of the original distribution of raw scores are duplicated in the distribution of the standard scores. Consequently, any computations that can be carried out with the original raw scores can also be carried out with the standard scores, without any distortion of results.

Why is it Necessary to Standardize Scores?

Standardizing scores is a mathematical means of ensuring that two or more scores, resulting from differing scales or units of measurement, have a common base and a common meaning and are therefore comparable. When used in the employment testing domain, standard scores ensure that scores from multiple components of a selection process are comparable even when, by design, the testing components have used differing scales of measurement. Standard scores also ensure that interview scores from two or more interview panels, using the same scale of measurement, are comparable.

When a selection process consisting of multiple testing components is developed, it is often assumed that the raw scores from each component are equivalent and may be merely combined to generate a final score which reflects the intended weight of each of the individual components. This is generally not true, however. For example, in the case of a selection process consisting of a 125-item multiple choice written examination (each item worth 1 point) and a performance test with a maximum score of 60 points, the written examination carries more than twice as much weight as the performance test.

Combining raw scores from different components of a selection process is similar to combining the measures of one inch and one foot without first converting them to a common base. A simple comparison of the numbers would indicate that they are the same value (i.e., 1 and 1), such that combining the two measures would yield a total of two. However, the individual measures and the corresponding total do not share a common unit of measurement. Consequently, it is unclear what the total really equals. Is it two inches? Is it two feet? If the two measures are converted to a common base, it can then be determined that one foot equals 12 inches. Thus, one foot is 12 times greater than one inch. This is a significant difference in the relative value of the two measures.

The same principles apply when combining candidate scores in a multiple hurdle selection process. In the example discussed above, a selection process was described which consists of a 125-item multiple choice written examination and a performance test with a maximum of 60 points possible. Let's assume that the written exam has a pass point of 92 points and the performance test has a pass point of 32 points. Let's further assume that each component is to be weighted 50% in the process – that is, a candidate's final score should be comprised of an equally weighted combination of his/her score on the written exam and his/her score on the performance test. If a candidate scores 98 on the written exam and 56 on the performance test, what is the candidate's final score in the process? It may seem completely acceptable to take the candidate's score on the written exam and simply add it to his/her score on the performance test without converting the component scores to a common base. If this approach is utilized, the candidate's final score would be 154 (i.e., 98 + 56). However, does a score of 154 accurately reflect the candidate's performance in the selection process? Furthermore, is each component in the process carrying an equal weight in determining the candidate's final score? The answer to both questions is a resounding "no." In this case, 64% of the candidate's final score would be derived from his/her performance on the written exam, and 36% would be derived from the performance test. Although the selection plan called for each of the two components to carry equal weight, in actuality this has not been accomplished.

Another situation in which there is a need to standardize scores is when a selection process for a particular job classification uses more than one interview panel. The assumption when using multiple interview panels is that each panel will apply the rating criteria in the same manner and that the rating process will be devoid of rating errors. This may not always be the case, however. Different interview panels may have different rating tendencies. One panel may be a very lenient panel and may tend to rate candidates towards the high end of the rating scale, while another panel may be more stringent and may tend to use the lower end of the scale.

Even when the interview process utilizes a highly structured approach (as should always be the case), it is not known for certain whether the two panels would rate a particular interview question response or a particular set of candidate qualifications in exactly the same manner. For example, assume that two candidates are interviewed for a job, each by a different interview panel. Candidate one receives a final interview score of 80, while candidate two receives a final interview score of 75. Although there is a score difference of five points between the two candidates, it is not known whether these results are due to actual differences in candidate qualifications – that is, candidate one is really five points better qualified than candidate two, or whether the results are due to differing applications of the rating criteria. Standardizing scores by interview panel, in essence, makes a mathematical correction for panel rating tendencies and assigns a common meaning to all interview scores which then allows for the comparison of scores between interview panels.

How are Standard Scores Calculated?

There are several different standard scores. The basic standard score is the z-score. Other standard scores are derived from the z-score. The z-score indicates the difference between a particular candidate's score and the mean score for the candidate group as a whole. The mean for z-scores is always 0.0 and the standard deviation is 1.0.

The z-score is calculated as follows:

$$z = \frac{X - \overline{X}}{S}$$

Where: X = a specific raw score
$\overline{X}$ = the mean raw score for the candidate group
S = the standard deviation for the candidate group

Example

- A candidate's raw score is 49
- The mean raw score for the candidate group is 40
- The standard deviation for the candidate group is 6

$$z = \frac{49 - 40}{6} = \frac{9}{6} = 1.5$$

The candidate's score is 1.5 standard deviations above the mean score for the candidate group.

One problem with z-scores is that they are given in both positive and negative values. Any scores below the mean are negative, while those above the mean are positive. An additional problem with z-scores is that they tend to produce decimals, as in the above example. Because of these characteristics, z-scores are not easy to work with and often lead to computational errors in those cases where they are calculated by hand. To overcome these difficulties, z-scores can quite easily be converted to T-scores.

The T-score is one of the most common linear standard scores. A T-score is derived by multiplying the z-score by 10 and adding 50 to the resulting product. T-scores have a mean of 50 and a standard deviation of 10.

$$T = 10z + 50$$

Where: $z = \frac{X - \overline{X}}{S}$

Example

- A candidate's raw score is 49
- The mean raw score for the candidate group is 40
- The standard deviation for the candidate group is 6

$$z = \frac{49 - 40}{6} = \frac{9}{6} = 1.5$$

$$T = 10(1.5) + 50 = 15 + 50 = 65$$

The candidate's T-score is 65. As with the z-score example above, the candidate's T-score is still 1.5 standard deviations above the mean for the candidate group as a whole.

In Summary:

Standardizing scores is merely a way to ensure that numbers that are being compared or combined have a common base and a common meaning. Standard scores have properties which make them more valuable and usable than raw scores:

- For every candidate group and exam, standard scores provide the same mean and the same standard deviation.
- Standard scores always retain the shape of the original raw score distribution.
- Standard scores permit intergroup and intertest comparisons that are not possible with raw scores.
- Standard scores can be treated mathematically. That is, they can be averaged and combined unlike raw scores.

Although standard scores are fairly easy to calculate by hand, there is rarely a need to do so. Many computer software packages, such as Microsoft Excel or Statistical Package for the Social Sciences (SPSS), can be used to generate the standard scores. Once the standard scores have been calculated, they can be easily weighted and combined as desired.

This monograph is intended to provide a general introduction to the use and application of standard scores. A more complete discussion of this topic can be found in most introductory statistics books.

Published by the Test Validation and Construction Unit
of the California State Personnel Board.
Visit our Webpage at www.spb.ca.gov/tvchome.htm.